

KEERTHIRAJAN SENTHILKUMAR

San Francisco, CA (Open to Relocate) | +1 (571) 220-6152 | keerthirajans@gwu.edu | LinkedIn | GitHub | Portfolio

Technical Skills

Languages: Python, TypeScript/JavaScript, SQL, C++, Bash

Full-Stack & APIs: React, Next.js (App Router), Streamlit, REST APIs, FastAPI, Django, Node.js

ML/AI: PyTorch, Hugging Face, LoRA/PEFT, Transformers, NLP, NumPy, Pandas, vector embeddings, model evaluation, LLM-as-judge

LLM Apps: LangChain, RAG, prompt engineering, agentic workflows, structured outputs, tool calling

Platform & Cloud: Docker, Git, CI/CD, pytest, Azure ML, AWS, Firebase, model monitoring, telemetry

Professional Experience

Apple

Seattle, WA

AI/ML Intern, Siri Core Modeling

May 2025 -- Aug 2025

- Built an LLM tool-calling evaluation pipeline on a 1,332-query dataset spanning 9 API domains with structured outputs.
- Fine-tuned Llama 3.1 8B with LoRA, raising accuracy from 60.9% to 70.9% & clarification accuracy from 80.5% to 95.1%.
- Wrote a validation suite for schema compliance, argument formatting, exact-match accuracy, and macro-F1 on every run.
- Packaged training and evaluation as reproducible scripts so teammates could rerun every experiment end to end.
- Automated LLM-as-judge scoring with the GPT-4o API, cutting manual review time 60% via regression pipelines.
- Benchmarked Gorilla OpenFunctions v2 (33.3% accuracy), isolating dispatch failures; presented to Siri AI/ML leadership.

TCS iON

Chennai, India

Machine Learning Intern

Nov 2021 -- Dec 2021

- Owned an NLP sentiment classifier end to end: trained the model, built a Flask UI, connected via a Django backend.
- Shipped a Dockerized service on live Twitter streams (Tweepy) with REST APIs: 95%+ uptime, 10K+ daily requests.
- Added automated monitoring and relevance metrics to the ingestion and ranking APIs, with a streaming retraining loop.
- Raised accuracy 18% over a Naive Bayes baseline (TF-IDF, n-grams), cut latency 40%, and added pytest checks for CI.

Research

When to Act and When to Ask: Tool Ambiguity in LLMs

Python, PyTorch, Hugging Face, PEFT GitHub

Continued after the internship on the open-source Berkeley Function Calling Leaderboard (BFCL) framework

- Fine-tuned Llama 3.1 8B with LoRA (rank 16, 0.52% of parameters): exact match 60.9% to 70.9%, 77.3% by LLM judge.
- Wrote a research paper draft with a senior Siri ML engineer as co-author, reported weaker tool-selection accuracy (56.5%) as a known limitation.
- Scored with exact match, a GPT-4o-mini judge, McNemar's test, bootstrap CIs; weakest area: tool selection, 56.5%.
- Trained in BF16 on a single 48GB A40 GPU in 1h 26m; published the code and documentation on GitHub.

Personal Projects

Historical Event Narrator - Fine-Tuning 7B LLM

Mistral-7B, PyTorch, LoRA/PEFT, Hugging Face, Streamlit

- Fine-tuned Mistral-7B with LoRA/PEFT: ROUGE-L up 100%+, quality 3/10 to 8/10 on a held-out benchmark.
- Synthesized 1,200+ training pairs from Wikipedia via Gemini 2.5 Flash; ran A/B comparisons with LLM-as-judge scoring.
- Led a 3-person team; built a Streamlit UI to switch between base and fine-tuned models for side-by-side comparison.
- Fixed mode collapse and hallucination with guided inference and stopping criteria, in a modular, tested codebase.

BarterBrAI - Multi-Modal AI Barter Platform

Gemini 2.5, Node.js, Firebase Cloud Functions, Flutter

- Built the AI backend as Node.js REST endpoints on Firebase Cloud Functions for valuation and negotiation coaching.
- Designed multimodal LLM valuation with confidence thresholds and a deterministic fallback for invalid LLM output.
- Integrated the AI endpoints into the Flutter UI; won Best Hack for Social Good at Technica 2025 (1st of 80+ teams).
- Wrote API docs, a Flutter integration guide, and Node.js test scripts so the app team could build against each endpoint.

Semantic Job-to-Problem Recommender

Sentence Transformers, FastAPI, Azure ML, LangChain, Streamlit

- Embedded 1,571 LeetCode problems (384-d Sentence Transformers) with two-stage retrieval, then LLM-based re-ranking.
- Deployed the ranker on Azure ML behind a FastAPI REST backend, with embeddings precomputed offline for speed.
- Built a Streamlit UI on the model that shows Grok's reasoning for each recommendation directly to users.

Education

The George Washington University

Washington, DC

M.S. in Computer Science, GPA: 3.84/4.00

Aug 2024 -- May 2026

Coursework: Machine Learning, Neural Networks, Distributed Systems, Cloud Computing, Data Mining, Software Paradigms

Sri Venkateswara College of Engineering

Chennai, India

B.E. in Computer Science, GPA: 3.54/4.00

Aug 2019 -- May 2023

Coursework: Data Structures, Machine Learning and Algorithms, Advanced UI Technologies, Artificial Intelligence

Achievements & Certifications

- Hackathon Winner:** BarterBrAI, Best Hack for Social Good, Technica 2025, 1st of 80+ teams and 450+ participants.
- AWS Academy Graduate:** Cloud Foundations, Cloud Operations, Machine Learning Foundations.
- Red Hat Certified System Administrator (RHCSA).**